Stochastic bounds for two-layer loss systems

M. Jonckheere* L. Leskelä†

February 15, 2008

Abstract

This paper studies multiclass loss systems with two layers of servers, where each server at the first layer is dedicated to a certain customer class, while the servers at the second layer can handle all customer classes. The routing of customers follows an overflow scheme, where arriving customers are preferentially directed to the first layer. Stochastic comparison and coupling techniques are developed for studying how the system is affected by packing of customers, altered service rates, and altered server configurations. This analysis leads to computationally fast upper and lower bounds for the performance of the system.

Keywords: multiclass loss system, overflow routing, maximum packing, stochastic order, preorder, coupling

AMS Subject Classification: 60K25, 60E15, 68M20, 90B15, 90B22

1 Introduction

This paper studies multiclass loss systems with two layers of servers, where each server at the first layer is dedicated to a certain customer class, and the servers at the second layer can handle all customer classes. Arriving customers are routed to vacant servers in one of the layers, with preference given to the first layer; or rejected otherwise. This policy is commonly referred to as overflow routing.

*Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands. http://homepages.cwi.nl/~jonckhee/

†Helsinki University of Technology, PO Box 1100, 02015 TKK, Finland. http://math.tkk.fi/~lleskela/

Layered networks with overflow routing are commonly used in telecommunications services, because different layers of service may increase the system capacity. In *wireless communication networks* for instance, the servers at the first layer correspond to radio channels dedicated to a small geographical area (microcell), and the second layer represents available radio channels in a larger area covering several microcells; in *telephone call centers*, the first layer consists of call agents trained to handling certain types of phone calls, and the second layer represents call agents who are cross-trained to deal with all types of calls.

The analysis of multilayer loss systems is challenging even under the simplest statistical assumptions, because the distributions of the overflow processes from the first layer are complex, and the direct numerical computation of the stationary distribution is unfeasible even for relatively small systems (Louth, Mitzenmacher, and Kelly [11]). Hence, approximative methods are needed for performance analysis (see Kelly [8] for a broad overview). Classical approximation techniques such as the equivalent random method and the Hayward–Fredericks method [18], and the recently introduced hyperexponential decomposition (Franx, Koole, and Pot [3]), are based on parametrically modeling the overflow processes from the first layer by simpler processes. These methods have been observed to produce good approximations for many choices of system parameters. However, they may require considerable amounts of computation, and it is not clear whether they remain accurate over the full parameter range.

The goal of this paper is to approximate the system via upper and lower bounds that are easy to compute numerically, and conservative in the sense that the true performance remains between the bounds for all choices of system parameters. To construct the upper bound, we modify the system by redirecting customers from the second layer into the first layer as soon as servers become vacant. This so-called maximum packing policy causes the number of customers per class to have a product-form stationary distribution (Everitt and Macfadyen [2]). The lower bound is constructed by moving all servers from the second layer into the first, this way reducing the system into a product of independent Erlang loss models.

The main tools for proving the validity of the bounds are (i) Massey's theorem [12] characterizing the comparability of two Markov jump processes; and (ii) stochastic coupling, where versions of the processes describing the number of customers in the original and the reference system are constructed in such a way that the difference of the two processes remains positive with probability one. Coupling techniques have been successfully used by several authors for deriving stochastic bounds for loss systems: Whitt [17] ana-

lyzed several single-class queueing systems; Smith and Whitt [15] studied the merging of two loss systems together; Nain [14] focused on multiclass single-layer loss systems; and Hordijk and Ridder [4] studied a special case of the two-layer loss system where the first layer is fully dedicated to a single customer class. This paper extends some of the above results to general multiclass two-layer loss systems, the main contribution being in showing that maximum packing leads to upper bounds for the time-dependent and stationary distributions of the number of customers in the system. In the special case where the first layer is fully dedicated to a single customer class, this result improves the upper bound obtained by Hordijk and Ridder [4].

The paper is organized as follows. Section 2 introduces the model details and notation. In Section 3 we prove a preliminary comparison result that is key to analyzing the monotonicity of the system. Section 4 analyzes how the time-dependent distribution of the system is affected by maximum packing, different server configurations, and altered service rates, and in Section 5 we carry out a similar analysis for the system in steady state. Section 6 concludes the paper.

2 Model description

2.1 Two-layer loss system with overflow routing

We consider a loss system with K customer classes and two layers of servers, where layer 1 contains m_k servers dedicated to class k, and layer 2 consists of n servers capable of serving all customer classes. Arriving class-k customers are routed to vacant servers in one of the layers, with preference given to layer 1; or rejected otherwise (Figure 1). For analytical tractability, we assume that the interarrival times and the service requirements of class-k customers are exponentially distributed with parameters λ_k and μ_k, respectively, and that all these random variables across all customer classes are independent. For brevity, we denote $m = (m_1, \ldots, m_K)$, $\lambda = (\lambda_1, \ldots, \lambda_K)$, and $\mu = (\mu_1, \ldots, \mu_K)$.

Denote by $X_{i,k}(t)$ the number of class-k customers being served at layer i at time t. The system is described by the continuous-time stochastic process $X = (X_{i,k})$ taking values in

$$S = \{x \in \mathbb{Z}_+^K \times \mathbb{Z}_+^K : \; x_{1,k} \leq m_k \; \forall k, \; \sum_{k=1}^{K} x_{2,k} \leq n\}. \tag{1}$$

Following the usual convention, we assume without loss of generality that

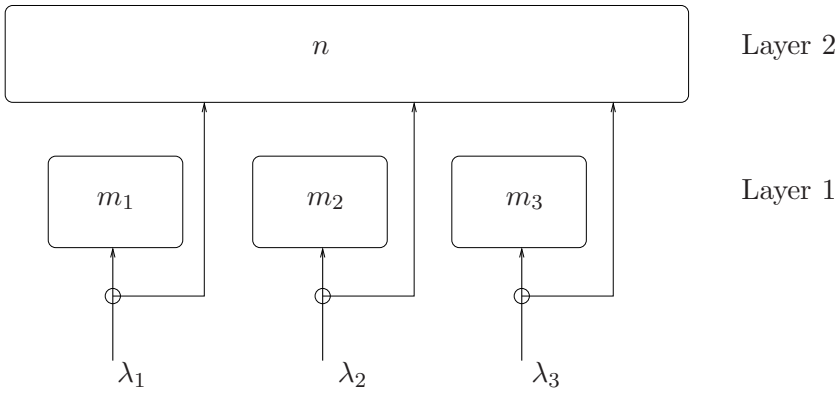

Figure 1: Two-layer loss network with three customer classes ($K = 3$).

all processes have paths in the space $D(\mathbb{R}_+, S)$ of right-continuous functions with left-hand limits [5].

Let us denote by $e_{i,k}$ the unit vector in $\mathbb{Z}_+^K \times \mathbb{Z}_+^K$ corresponding to the coordinate direction (i, k). Moreover, define the sets

$$A_{1,k} = \{x \in S : x_{1,k} < m_k\}, \tag{2}$$

$$A_{2,k} = \{x \in S : x_{1,k} = m_k,\ \sum_{l=1}^{K} x_{2,l} < n\}, \tag{3}$$

$$B_k = \{x \in S : x_{1,k} = m_k,\ \sum_{l=1}^{K} x_{2,l} = n\}. \tag{4}$$

The set $A_{i,k}$ represents the set of states where an arriving class-k customer is assigned to a layer-i server, and B_k is the set of states where arriving class-k customers are rejected. The process X is a continuous-time Markov process on S with the upward transitions $x \mapsto x + e_{i,k}$ at rate $\lambda_{i,k}(x)$, and downward transitions $x \mapsto x - e_{i,k}$ at rate $\phi_{i,k}(x)$, where

$$\begin{aligned} \lambda_{i,k}(x) &= \lambda_k 1(x \in A_{i,k}), \\ \phi_{i,k}(x) &= \mu_k x_{i,k}. \end{aligned} \tag{5}$$

2.2 Maximum packing

To approximate the original two-layer loss system, we consider a modification of the system, where customers are redirected from layer 2 to layer 1 as soon as servers become vacant. This corresponds to the so-called maximum packing policy introduced by Everitt and Macfadyen [2]. The process

X^{mp} describing the number of customers in this system is a continuous-time Markov process on S with the upward transitions $x \mapsto x+e_{i,k}$ at rate $\lambda'_{i,k}(x)$, and downward transitions $x \mapsto x - e_{i,k}$ at rate $\phi'_{i,k}(x)$, where

$$\begin{aligned} \lambda'_{i,k}(x) &= \lambda_k 1(x \in A_{i,k}), \quad i = 1, 2, \\ \phi'_{1,k}(x) &= \mu_k x_{1,k} 1(x_{2,k} = 0), \\ \phi'_{2,k}(x) &= \mu_k x_{1,k} 1(x_{2,k} > 0) + \mu_k x_{2,k}. \end{aligned} \tag{6}$$

Remark 1. A remarkable property of the maximum packing policy is that all states outside the set $S^{\mathrm{mp}} = \cap_{k=1}^{K} \{x \in S : x_{1,k} = m_k \text{ or } x_{2,k} = 0\}$ are transient for X^{mp}. Moreover, note that for $x \in S^{\mathrm{mp}}$, $x_{1,k} = m_k$ if and only if $x_{1,k} + x_{2,k} \geq m_k$, which implies that

$$\begin{aligned} x_{1,k} &= (x_{1,k} + x_{2,k}) \wedge m_k, \\ x_{2,k} &= (x_{1,k} + x_{2,k} - m_k)^+. \end{aligned} \tag{7}$$

As a consequence, the aggregate process $(X^{\mathrm{mp}}_{1,k} + X^{\mathrm{mp}}_{2,k})_{k=1}^{K}$ tracking the total number of customers in each class, if started in S^{mp}, is equal in distribution to the Markov process $\hat{X}^{\mathrm{mp}} = (\hat{X}^{\mathrm{mp}}_1, \ldots, \hat{X}^{\mathrm{mp}}_K)$ on $\hat{S}^{\mathrm{mp}} = \{\hat{x} \in \mathbb{Z}_+^K : \sum_k (\hat{x}_k - m_k)^+ \leq n\}$ generated by the transitions

$$\hat{x} \mapsto \begin{cases} \hat{x} + e_k, & \text{at rate } \lambda_k 1(\hat{x} + e_k \in \hat{S}^{\mathrm{mp}}), \\ \hat{x} - e_k, & \text{at rate } \phi_k(\hat{x}) = \mu_k \hat{x}_k. \end{cases}$$

The structure of the above transition rates implies that the stationary distribution of $\hat{X}^{\mathrm{mp}}$ is a product of Poisson distributions truncated to $\hat{S}^{\mathrm{mp}}$ [8], which is easy to compute numerically. The stationary distribution of X^{mp} can then be recovered from that of $\hat{X}^{\mathrm{mp}}$ using the equalities (7).

3 Preliminary result

This section establishes a general result that allows to compare two processes taking values in $S \subset \mathbb{Z}_+^K \times \mathbb{Z}_+^K$ with respect to a specific preorder. This preorder, tailored to fit the transition rates of the type in (5), is defined by $x \preceq y$, if $x_{1,k} \leq y_{1,k}$ for all k and $|x| \leq |y|$, where $|x| = \sum_{i,k} x_{i,k}$. For random variables with values in S we denote $U \preceq_{\mathrm{st}} V$, if $\mathrm{E}\,\phi(U) \leq \mathrm{E}\,\phi(V)$ for all bounded measurable functions $\phi : S \to \mathbb{R}$ that are increasing with respect to the preorder $\preceq$ on S. Let us further extend these definitions to the Skorohod space $D(\mathbb{R}_+, S)$ of right-continuous functions with left-hand limits by denoting $f \preceq g$ if $f(t) \preceq g(t)$ for all t. For stochastic processes with

paths in $D(\mathbb{R}_+, S)$ we denote $X \preceq_{\mathrm{st}} Y$, if $\mathrm{E}\,\phi(X) \le \mathrm{E}\,\phi(Y)$ for all bounded measurable maps $\phi : D(\mathbb{R}_+, S) \to \mathbb{R}$ that are increasing with respect to the preorder $\preceq$ on $D(\mathbb{R}_+, S)$. It will be clear from the context whether $\preceq$ refers to elements in S or to functions in $D(\mathbb{R}_+, S)$.

Consider a continuous-time Markov process X on $S \subset \mathbb{Z}_+^K \times \mathbb{Z}_+^K$ generated by the transitions

$$x \mapsto \begin{cases} x + e_{i,k} & \text{at rate } \lambda_{i,k}(x), \\ x - e_{i,k} & \text{at rate } \phi_{i,k}(x), \end{cases}$$

$i \in \{1, 2\}$, $k \in \{1, \dots, K\}$, where $\lambda_{i,k}$ and $\phi_{i,k}$ are bounded nonnegative functions on S. For consistency, we assume here that $\lambda_{i,k}(x) = 0$ for all $x \in S$ such that $x + e_{i,k} \notin S$ and $\phi_{i,k}(x) = 0$ for all $x \in S$ such that $x - e_{i,k} \notin S$. We assume that Y is a similar process with state-dependent transition rates $\lambda'_{i,k}$ and $\phi'_{i,k}$.

Theorem 1. *Let X and Y be Markov processes with paths in $D(\mathbb{R}_+, S)$ having upward transition rates $\lambda_{i,k}$ and $\lambda'_{i,k}$, and downward transition rates $\phi_{i,k}$ and $\phi'_{i,k}$, respectively. Assume that the following two conditions hold:*

(i) For all $x, y \in S$ such that $x \preceq y$ and $x_{1,k} = y_{1,k}$,

$$\lambda_{1,k}(x) \le \lambda'_{1,k}(y), \tag{8}$$

$$\phi_{1,k}(x) \ge \phi'_{1,k}(y). \tag{9}$$

(ii) For all $x, y \in S$ such that $x \preceq y$ and $|x| = |y|$,

$$\sum_{i,k} \lambda_{i,k}(x) \le \sum_{i,k} \lambda'_{i,k}(y), \tag{10}$$

$$\sum_{i,k} \phi_{i,k}(x) \ge \sum_{i,k} \phi'_{i,k}(y). \tag{11}$$

Then $X \preceq_{\mathrm{st}} Y$, given that the initial states satisfy $X(0) \preceq Y(0)$.

Proof. Denote the infinitesimal generators of X and Y by p and q, respectively. Recall that $U \subset S$ is called an *upper set*, if $x \in U$ and $x \preceq y$ implies $y \in U$, and $V \subset S$ is called a *lower set*, if the complement V^c of V is an upper set. Using a result of Massey [12, Theorem 5.3][1] (see also [6, Theorem

[1] Massey formulated his result for partially ordered spaces, but all the proofs in his paper [12] remain valid also for preorders that are not antisymmetric [9].

5]), it suffices to verify that $p(x,U) \le q(y,U)$ for all $x \preceq y$ and for all upper sets U such that either $x \in U$ or $y \notin U$. Because $p(x,U) = -p(x,U^c)$ for all $x \in U$, this condition is equivalent to showing that for all $x \preceq y$,

$$\sum_{i,k} \lambda_{i,k}(x)\, 1(x + e_{i,k} \in U) \le \sum_{i,k} \lambda'_{i,k}(y)\, 1(y + e_{i,k} \in U) \tag{12}$$

for all upper sets U such that $x \notin U, y \notin U$, and

$$\sum_{i,k} \phi_{i,k}(x)\, 1(x + e_{i,k} \in V) \ge \sum_{i,k} \phi'_{i,k}(y)\, 1(y + e_{i,k} \in V) \tag{13}$$

for all lower sets V such that $x \notin V, y \notin V$.

Assume $x \preceq y$ and choose an upper set U such that $x \notin U, y \notin U$. To verify the validity of (12), let us consider separately the cases $|x| < |y|$ and $|x| = |y|$. Assume first $|x| < |y|$. Then $x + e_{1,k} \preceq y$ for all k such that $x_{1,k} < y_{1,k}$, and $x + e_{2,k} \preceq y$ for all k. Hence because U is an upper set and $y \notin U$, it follows that $x + e_{1,k} \in U$ only if $x_{1,k} = y_{1,k}$, and $x + e_{2,k} \notin U$ for all k. Thus,

$$\sum_{i,k} \lambda_{i,k}(x)\, 1(x + e_{i,k} \in U) = \sum_{k : x_{1,k} = y_{1,k}} \lambda_{1,k}(x)\, 1(x + e_{1,k} \in U). \tag{14}$$

Moreover, using inequality (8), and noting that $x + e_{1,k} \preceq y + e_{1,k}$ for all k such that $y + e_{1,k} \in S$, we see that for all k such that $x_{1,k} = y_{1,k}$,

$$\lambda_{1,k}(x)\, 1(x + e_{1,k} \in U) \le \lambda'_{1,k}(y)\, 1(y + e_{1,k} \in U). \tag{15}$$

Substituting (15) into (14) shows the validity of (12).

Let us next focus on the case $|x| = |y|$. Note first that if $x + e_{1,l} \in U$ for some l such that $x_{1,l} < y_{1,l}$, or $x + e_{2,l} \in U$ for some l, then $y + e_{i,k} \in U$ for all i and k. Hence it follows that the right-hand side of (12) equals $\sum_{i,k} \lambda'_{i,k}(y)$, which in light of assumption (10) guarantees the validity of (12). On the other hand, if $x + e_{2,k} \notin U$ for all k, and $x_{1,k} = y_{1,k}$ for all k such that $x + e_{1,k} \in U$, then equation (14) holds. Assumption (8) again implies (15), which together with (14) shows the validity of (12).

The proof is completed by carrying out an analogous reasoning for lower sets, which shows that assumptions (9) and (11) imply (13). □

4 Pathwise stochastic comparison

This section contains the main results for analyzing the time-dependent distribution of the system. Assuming first that all service rates across different

customer classes are equal, we study how the system is affected by maximum packing (Section 4.1) and different server configurations (Section 4.2). Section 4.3 provides a monotonicity result that allows to extend the analysis to the case where the service rates are not assumed equal, and Section 4.4 describes bounds for the per-class number of customers in the system.

Recall that the usual stochastic order [13] between real random variables is defined by denoting $U \leq_{\mathrm{st}} V$, if $\mathrm{E}\, f(U) \leq \mathrm{E}\, f(V)$ for all bounded measurable increasing real functions f. Moreover, for stochastic processes with paths in the Skorohod space $D(\mathbb{R}_+, \mathbb{R})$, we denote $X \leq_{\mathrm{st}} Y$ if $\mathrm{E}\, f(X) \leq \mathrm{E}\, f(Y)$ for all bounded measurable functions $f : D(\mathbb{R}_+, \mathbb{R}) \to \mathbb{R}$ that are increasing with respect to the natural pointwise order on $D(\mathbb{R}_+, \mathbb{R})$. A *coupling* of two stochastic processes X and Y with paths in $D(\mathbb{R}_+, \mathbb{R})$ is a stochastic process $(\hat{X}, \hat{Y})$ with paths in $D(\mathbb{R}_+, \mathbb{R}^2)$, having X and Y as its marginals. Recall that by Strassen's theorem, $X \leq_{\mathrm{st}} Y$ if and only if there exists a coupling $(\hat{X}, \hat{Y})$ of X and Y such that $\hat{X}(t) \leq \hat{Y}(t)$ for all t almost surely [6]. Strassen's theorem can further be extended to processes with paths in $D(\mathbb{R}_+, S)$, compared with respect to a given preorder [10].

4.1 Maximum packing

Let X be the process describing the number of customers in the two-layer loss system defined in Section 2.1, and denote by X^{mp} the corresponding process for the maximum packing policy defined in Section 2.2. Recall from Section 3 that the preorder $x \preceq y$ is defined by $x_{1,k} \leq y_{1,k}$ for all k and $|x| \leq |y|$. The following theorem is the main result of the paper. It allows to conclude that the stochastic processes $t \mapsto |X(t)|$ and $t \mapsto |X^{\mathrm{mp}}(t)|$ satisfy $|X| \leq_{\mathrm{st}} |X^{\mathrm{mp}}|$, given that $X(0) \preceq X^{\mathrm{mp}}(0)$.

Theorem 2. *Assume that all service rates μ_k are equal and that the initial states satisfy $X(0) \preceq X^{\mathrm{mp}}(0)$. Then $X \preceq_{\mathrm{st}} X^{\mathrm{mp}}$.*

Example 1 below shows that a purely deterministic sample path comparison is not sufficient for proving Theorem 2; hence probabilistic techniques are needed. Example 3 in Section 5.2 further shows that the statement of Theorem 2 may not be true, if the service rates are not assumed equal.

Example 1. Consider a two-class system ($K = 2$) with one server at layer 1 assigned to class 1 ($m_1 = 1, m_2 = 0$) and one server at layer 2 ($n = 1$). Denote by $X = (X_{i,k})$ a path of the process tracking the number of customers in the original two-layer loss system, and let X^{mp} be a corresponding sample path for the maximum packing policy. Assume that during the time interval

$[0, 6]$ there are four arriving customers each having service time equal to three: three class-1 arrivals at time epochs 0, 2, and 4; and one class-2 arrival at time epoch 3. Given that both systems start empty, then $X(6) = e_{1,1}$ but $X^{\mathrm{mp}}(6) = 0$.

Lemma 1. *The transition rates $\lambda_{i,k}(x)$ defined in (5) satisfy:*

(i) For all $x \preceq y$ and for all k such that $x_{1,k} = y_{1,k}$,

$$\lambda_{1,k}(x) \le \lambda_{1,k}(y). \tag{16}$$

(ii) For all $x \preceq y$ and for all k such that $|x| = |y|$,

$$\sum_{i,k} \lambda_{i,k}(x) \le \sum_{i,k} \lambda_{i,k}(y). \tag{17}$$

Proof. The inequality (16) is clear, because $\lambda_{1,k}(x)$ only depends on $x_{1,k}$. Assume next that $x \preceq y$ and $|x| = |y|$. Assume that $y \in B_k$ for some k, where B_k is defined in (4). Then $\sum_l y_{2,l} = n$, which implies that $\sum_l x_{2,l} = n$ and $x_{1,l} = y_{1,l}$ for all l. Thus $x \in B_k$. We may thus conclude that for all k, $1(x \notin B_k) \le 1(y \notin B_k)$. Hence it follows that

$$\sum_{i,k} \lambda_{i,k}(x) = \sum_k \lambda_k 1(x \notin B_k) \le \sum_k \lambda_k 1(y \notin B_k) = \sum_{i,k} \lambda_{i,k}(y),$$

which shows the validity of (17). □

Proof of Theorem 2. Let $\lambda_{i,k}(x)$ and $\phi_{i,k}(x)$ be the transition rates of X as defined in (5), and let $\lambda'_{i,k}(x)$ and $\phi'_{i,k}(x)$ be the corresponding rates for X^{mp} as defined in (6). Because $\lambda'_{i,k}(x) = \lambda_{i,k}(x)$ for all x, the validity of (8) and (10) in Theorem 1 follow by Lemma 1. For the downward transitions, note that for all $x \preceq y$ such that $x_{1,k} = y_{1,k}$ for some k, $\phi_{1,k}(x) = \mu_1 x_{1,k} = \mu_1 y_{1,k} \ge \mu_1 y_{1,k} 1(y_{2,k} = 0) = \phi'_{1,k}(y)$. Moreover, for all $x \preceq y$ such that $|x| = |y|$,

$$\sum_k (\phi_{1,k}(x) + \phi_{2,k}(x)) = \mu_1 |x| = \mu_1 |y| = \sum_k (\phi'_{1,k}(y) + \phi'_{2,k}(y)),$$

so conditions (9) and (11) of Theorem 1 are valid. Hence Theorem 1 yields the claim. □

4.2 Different server configurations

This section studies the effect of moving one server from layer 1 to layer 2. As in Section 2.1, we denote by X the process describing the number of customers in the system with server configuration $m = (m_1, \dots, m_K)$ in layer 1, and n servers in layer 2. Let Y by the process corresponding to the modified system where one class-k server from layer 1 has been replaced by a server in layer 2. We assume $k = 1$ without loss of generality. Let $m' = (m_1 - 1, m_2, \dots, m_K)$ and $n' = n + 1$, and define the sets S', $A'_{1,k}$ and B'_k as in (1)–(4) with m and n replaced by m' and n', respectively. Then Y is a Markov process on S' having transition rates of the form (5) with $A_{i,k}$ replaced by $A'_{i,k}$.

Let us denote by $x_2 = \sum_k x_{2,k}$ the number of customers being served at layer 2. Assuming that all service rates μ_k are equal, it follows that the process $(X_{1,1}, \dots, X_{1,K}; X_2)$ is Markov. With a slight abuse of notation, we will redefine the state space by $S = \{(x_{1,1}, \dots, x_{1,K}; x_2) \in \mathbb{Z}_+^K \times \mathbb{Z}_+ : x_{1,k} \le m_k \text{ for all } k, x_2 \le n\}$, and denote by e_2 the unit vector in $\mathbb{Z}_+^K \times \mathbb{Z}_+$ corresponding to the last coordinate. We will redefine the sets $A_{i,k}, B_k, A'_{i,k}, B'_k$, and S' in a similar way, identifying $\sum_{k=1}^K x_{2,k}$ with x_2.

Theorem 3. *Assume that all service rates μ_k are equal, and that the initial states satisfy $Y(0) - X(0) \in \Delta$, where $\Delta = \{0,\, e_2,\, e_2 - e_{1,1},\, 2e_2 - e_{1,1}\}$. Then the stochastic processes $t \mapsto |X(t)|$ and $t \mapsto |Y(t)|$ satisfy $|X| \le_{\mathrm{st}} |Y|$.*

Proof. Because $|x| \le |y|$ for all $x \in S$ and $y \in S'$ such that $y - x \in \Delta$, it is sufficient to construct a coupling [16] of X and Y that takes values in $S_\Delta = \{(x, y) \in S \times S' : y - x \in \Delta\}$. Let $(\tilde{X}, \tilde{Y})$ be a continuous-time Markov process on S_Δ generated by the joint arrivals

$$
\begin{aligned}
&(x, y) \mapsto (x + e_{1,k},\, y + e_{1,k}) && \text{at rate } \lambda_k 1(x \in A_{1,k},\, y \in A'_{1,k}), && (18)\\
&(x, y) \mapsto (x + e_{1,k},\, y + e_2) && \text{at rate } \lambda_k 1(x \in A_{1,k},\, y \in A'_{2,k}), && (19)\\
&(x, y) \mapsto (x + e_{1,k},\, y) && \text{at rate } \lambda_k 1(x \in A_{1,k},\, y \in B'_k), && (20)\\
&(x, y) \mapsto (x + e_2,\, y + e_2) && \text{at rate } \textstyle\sum_l \lambda_l 1(x \in A_{2,l},\, y \in A'_{2,l}), && (21)\\
&(x, y) \mapsto (x + e_2,\, y) && \text{at rate } \textstyle\sum_l \lambda_l 1(x \in A_{2,l},\, y \in B'_l), && (22)\\
&(x, y) \mapsto (x,\, y + e_2) && \text{at rate } \textstyle\sum_l \lambda_l 1(x \in B_l,\, y \in A'_{2,l}), && (23)
\end{aligned}
$$

and joint departures

$$(x,y) \mapsto (x - e_{1,k},\, y - e_{1,k}) \quad \text{at rate} \quad \mu_1 y_{1,k}, \tag{24}$$
$$(x,y) \mapsto (x - e_{1,1},\, y - e_2) \quad \text{at rate} \quad \mu_1 (x_{1,1} - y_{1,1}), \tag{25}$$
$$(x,y) \mapsto (x - e_2,\, y - e_2) \quad \text{at rate} \quad \mu_1 x_2, \tag{26}$$
$$(x,y) \mapsto (x,\, y - e_2) \quad \text{at rate} \quad \mu_1 (y_{1,1} + y_2 - x_{1,1} - x_2). \tag{27}$$

Observe that all transition rates above are nonnegative, because $y_{1,1} \leq x_{1,1}$ and $y_{1,1} + y_2 \geq x_{1,1} + x_2$, whenever $y - x \in \Delta$. To ensure that the transitions define a generator of a Markov process on S_Δ, we need to verify that $y' - x' \in \Delta$ for all transitions $(x,y) \mapsto (x',y')$, where $y - x \in \Delta$. This is obvious for transitions (18), (21), (24), and (26), because in these cases $y' - x' = y - x$. Let us consider the remaining cases one-by-one:

- If transition (19) occurs, then $k = 1$, because $y_{1,k} = x_{1,k}$ for all $k \neq 1$. Then $x_{1,1} < m_1$ and $y_{1,1} = m_1 - 1$, so it follows that either $y - x = 0$ or $y - x = e_2$. In both cases, $y' - x' \in \Delta$.

- If transition (20) occurs, then again $k = 1$. Then $x_{1,1} < m_1$ and $y_{1,1} = m_1 - 1$, which implies $y_{1,1} = x_{1,1}$. Moreover, $y_2 = n + 1$, which is only possible if $y_2 = x_2 + 1$. Hence $y - x = e_2$, so that $y' - x' = e_2 - e_{1,1} \in \Delta$.

- If transition (22) occurs, then $x \in A_{2,l}$ and $y \in B'_l$ for some l. Then $x_2 < n$ and $y_2 = n + 1$, which implies that $y - x = 2e_2 - e_{1,1}$. Hence $y' - x' = e_2 - e_{1,1} \in \Delta$.

- If transition (23) occurs, then $x \in B_l$ and $y \in A'_{2,l}$ for some l. Then $x_2 = n$ and $y_2 < n + 1$, so it follows that $y_2 = x_2$. Hence $y - x = 0$, and thus $y' - x' = e_2 \in \Delta$.

- If transition (25) occurs, then $y_{1,1} < x_{1,1}$. Because $y - x \in \Delta$, this implies that either $y - x = e_2 - e_{1,1}$, so that $y' - x' = 0$; or $y - x = 2e_2 - e_{1,1}$, so that $y' - x' = e_2$.

- If transition (27) occurs, then $y_{1,1} + y_2 - x_{1,1} - x_2 > 0$. Because $y - x \in \Delta$, it follows that either $y - x = e_2$, so that $y' - x' = 0$; or $y - x = 2e_2 - e_{1,1}$, so $y' - x' = e_2 - e_{1,1}$.

Hence, all transitions map S_Δ into S_Δ, and the process $(\tilde{X}, \tilde{Y})$ is well-defined.

To show that $(\tilde{X}, \tilde{Y})$ is a coupling of X and Y, we must verify that the marginal transition rates of $(\tilde{X}, \tilde{Y})$ match with the transition rates of X and

Y. Note first that the sum of transition rates such that $x \mapsto x + e_{1,k}$ is equal to $\lambda_k 1(x \in A_{1,k})$. Next, observe that $x \in A_{2,l}$ and $y - x \in \Delta$ imply that $y \notin A'_{1,l}$. Hence the sum of transition rates where $x \mapsto x + e_2$ is equal to

$$\sum_l \lambda_l 1(x \in A_{2,l}, y \in A'_{2,l} \cup B'_l) = \sum_l \lambda_l 1(x \in A_{2,l}).$$

Further, because the sum of all transition rates such that $x \mapsto x - e_{1,k}$ equals $\mu_1 x_{1,k}$ for all k, and the corresponding sum for $x \mapsto x - e_2$ is equal to $\mu_1 x_2$, we may conclude that the transitions of $\tilde{X}$ and X occur at the same rates.

Turning the attention to the rates of $\tilde{Y}$, note that $y - x \in \Delta$ and $y \in A'_{1,1}$ imply that $y_{1,1} < m_1 - 1$ and $x_{1,1} \leq y_{1,1} + 1$, so it follows that $x \in A_{1,1}$. Moreover, $y - x \in \Delta$ and $y \in A'_{1,k}$ for $k \neq 1$ imply that $x_{1,k} = y_{1,k} < m_k$, so $x \in A_{1,k}$. Hence the total rate of transitions where $y \mapsto y + e_{1,k}$ is equal to $\lambda_k 1(x \in A_{1,k},\, y \in A'_{1,k}) = \lambda_k 1(y \in A'_{1,k})$. Further, because the net rate of transitions where $y \mapsto y + e_2$ is equal to $\sum_l \lambda_l 1(y \in A'_{2,l})$, and because the corresponding net rates for $y \mapsto y - e_{1,k}$ and $y \mapsto y - e_2$ are equal to $\mu_1 y_{1,k}$ and $\mu_1 y_2$, respectively, we conclude that the transitions of $\tilde{Y}$ and Y occur at the same rates. Hence, the process $(\tilde{X}, \tilde{Y})$ is a coupling of X and Y. □

4.3 Monotonicity with respect to service rates

The results in Sections 4.1 and 4.2 were proved under the assumption that all service rates are equal. The following theorem describes a monotonicity property that allows to compare systems not satisfying this assumption. Denote by X the number of customers of the two-layer loss system defined in Section 2.1. Recall that the preorder $x \preceq y$ is defined by $x_{1,k} \leq y_{1,k}$ for all k and $|x| \leq |y|$.

Theorem 4. *Let X^- and X^+ be modifications of the system with all service rates set to $\mu_{\max} = \max \mu_k$ and $\mu_{\min} = \min \mu_k$, respectively. Assume that the initial states satisfy $X^-(0) \preceq X(0) \preceq X^+(0)$. Then*

$$X^- \preceq_{\mathrm{st}} X \preceq_{\mathrm{st}} X^+.$$

Remark 2. A simpler comparison statement, such as $|X| \leq_{\mathrm{st}} |X^+|$ given that $|X(0)| \leq |X^+(0)|$, is not true in general. Using Massey's [12] criteria for the preorder $|x| \leq |y|$, it is not hard to check that a necessary condition for the above property is that $\sum_{i,k} \lambda_{i,k}(x) = \sum_{i,k} \lambda_{i,k}(y)$ whenever $|x| = |y|$. This equality fails for $x = \sum_k m_k e_{1,k} + (n-1)e_{2,1}$ and $y = x - e_{1,1} + e_{2,1}$.

Proof of Theorem 4. Note that X^+ has the same upward transitions as X and downward transitions $\phi'_{1,k}(x) = \mu_{\min} x_{1,k}$, and $\phi'_{2,k}(x) = \mu_{\min} x_{2,k}$. Now for all $x \preceq y$ such that $x_{1,k} = y_{1,k}$ for some k, $\mu_k x_{1,k} \geq \mu_{\min} x_{1,k} = \mu_{\min} y_{1,k}$, and for all $x \preceq y$ such that $|x| = |y|$,

$$\sum_k \mu_k (x_{1,k} + x_{2,k}) \geq \mu_{\min} \sum_k (x_{1,k} + x_{2,k}) = \mu_{\min} \sum_{i,k} (y_{1,k} + y_{2,k}),$$

so conditions (9) and (11) of Theorem 1 are valid. Moreover, (8) and (10) hold by Lemma 1, so Theorem 1 yields the claim for X^+. The claim for X^- is proved in a similar way. □

4.4 Per-class bounds

In this section, we prove upper and lower bounds for the per-class number of customers in the system. Let $Z^s_{\lambda,\mu}(t)$ be the number of customers in the standard s-server Erlang loss system at time t, defined as the right-continuous Markov process on $\{0, 1, \ldots, s\}$ having the upward transitions $x \mapsto x + 1$ at rate $\lambda 1(x < s)$ and the downward transitions $x \mapsto x - 1$ at rate μx.

Theorem 5. *Assume* $Z^{m_k}_{\lambda_k,\mu_k}(0) = X_{1,k}(0)$. *Then*

$$Z^{m_k}_{\lambda_k,\mu_k} \leq_{\text{st}} X_{1,k} + X_{2,k}. \tag{28}$$

Proof. Observe that the process $X_{1,k}$ tracking the number of class-k customers being served at layer 1 has the same dynamics as a standard m_k-server Erlang loss system with arrival rate λ_k and service rate μ_k. Hence given $Z^{m_k}_{\lambda_k,\mu_k}(0) = X_{1,k}(0)$, the processes $Z^{m_k}_{\lambda_k,\mu_k}$ and $X_{1,k}$ have the same distribution, which immediately implies (28). □

Theorem 6. *Assume* $X_{1,k}(0) + X_{2,k}(0) \leq Z^{m_k+n}_{\lambda_k,\mu_k}(0)$. *Then*

$$X_{1,k} + X_{2,k} \leq_{\text{st}} Z^{m_k+n}_{\lambda_k,\mu_k}. \tag{29}$$

Proof. Assume without loss of generality that $k = 1$. Let us construct a Markov process $(\tilde{X}, \tilde{Y})$ on

$$S_2 = \{(x, y) \in S \times \{0, \ldots, m_1 + n\} : x_{1,1} + x_{2,1} \leq y\}$$

via the class-1 transitions for $i = 1, 2$,

$$(x, y) \mapsto (x + e_{i,1},\, y + 1) \quad \text{at rate} \quad \lambda_1 1(x \in A_{i,1},\, y < m_1 + n), \tag{30}$$
$$(x, y) \mapsto (x + e_{i,1},\, y) \quad \text{at rate} \quad \lambda_1 1(x \in A_{i,1},\, y = m_1 + n), \tag{31}$$
$$(x, y) \mapsto (x,\, y + 1) \quad \text{at rate} \quad \lambda_1 1(x \in B_1,\, y < m_1 + n), \tag{32}$$
$$(x, y) \mapsto (x - e_{i,1},\, y - 1) \quad \text{at rate} \quad \mu_1 x_{i,1}, \tag{33}$$
$$(x, y) \mapsto (x,\, y - 1) \quad \text{at rate} \quad \mu_1 (y - x_{1,1} - x_{2,1}), \tag{34}$$

and the class-k transitions for $k \neq 1$ and $i = 1, 2$,

$$(x, y) \mapsto (x + e_{i,k},\, y) \quad \text{at rate} \quad \lambda_k 1(x \in A_{i,k}), \tag{35}$$
$$(x, y) \mapsto (x - e_{i,k},\, y) \quad \text{at rate} \quad \mu_k x_{i,k}. \tag{36}$$

Note that all transition rates in (30) – (36) are nonnegative for all $(x, y) \in S_2$.

Let us now verify that all transitions map S_2 into S_2. Observe first that transition (31) occurs only if $y = m_1 + n$ and either $x_{1,1} < m_1$ or $\sum_{k=1}^{K} x_{2,k} < n$, which implies that $(x + e_{i,1},\, y) \in S_2$ for $i = 1, 2$. Moreover, transition (34) occurs only if $x_{1,1} + x_{2,1} < y$, so that $(x, y - 1) \in S_2$. Clearly, all other transitions map S_2 into S_2. Thus the Markov process $(\tilde{X}, \tilde{Y})$ on S_2 is well-defined.

Moreover, the total rates of transitions in (30) – (36) where $x \mapsto x + e_{i,k}$ and $x \mapsto x - e_{i,k}$ are equal to $\lambda_k 1(x \in A_{i,k})$ and $\mu_k x_{i,k}$, respectively, for all i and k. The corresponding total rates for $y \mapsto y + 1$ and $y \mapsto y - 1$ are equal to $\lambda_1 1(y < m_1 + n)$ and $\mu_1 y$, respectively. This shows that $(\tilde{X}, \tilde{Y})$ is a coupling of X and $Z_{\lambda_1, \mu_1}^{m_1+n}$, so the inequality (29) holds. □

Remark 3. The proof of Theorem 5 actually shows that inequality (28) can be extended to arbitrary (random or nonrandom) arrival processes and service times. Example 2 shows why this kind of purely deterministic sample path comparison is not possible for obtaining the result in Theorem 6.

Example 2. Consider a two-class system ($K = 2$) with no servers at layer 1 ($m_1 = 0, m_2 = 0$) and one server at layer 2 ($n = 1$). Denote by $X = (X_{i,k})$ a path of the process tracking the number of customers in the original two-layer loss system, and let Z be a corresponding sample path of the modified (one-class) system that only accepts class-1 customers. Assume that during the time interval $[0, 3]$ there are three arriving customers each having service time equal to two: a class-2 arrival at time epoch 0, and two class-1 arrivals at time epochs 1 and 2. Given that both systems start empty, then $X_{2,1}(3) = 1$ but $Z(3) = 0$.

5 Bounds of the steady-state performance

Assume from now on that all arrival rates and service rates are strictly positive, which implies that all Markov processes treated in the sequel have a unique stationary distribution. In this section $\bar{X} = (\bar{X}_{i,k})$ denotes a random vector describing the stationary number of class-k customers being served at layer i in the system, and the quadruple (m, n, λ, μ) indicates that a performance quantity corresponds to a system with server configuration $m = (m_1, \ldots, m_K)$ at layer 1, n servers at layer 2, arrival rates $\lambda = (\lambda_1, \ldots, \lambda_K)$, and service rates $\mu = (\mu_1, \ldots, \mu_K)$.

5.1 Per-class performance

Denote by $a_k = \mathrm{E}(\bar{X}_{1,k} + \bar{X}_{2,k})$ the stationary mean number of class-k customers in the system, by θ_k the stationary mean class-k throughput (the number of class-k customers completing service per unit time), and by b_k the class-k blocking probability. Note that a_k can be viewed as the mean class-k work throughput (amount of class-k work served per unit time).

Let $\mathrm{Erl}(s, \rho)$ be a random variable on $\{0, 1, \ldots, s\}$ having distribution $(\sum_{j=0}^{s} \frac{\rho^j}{j!})^{-1} \frac{\rho^i}{i!}$, and denote its mean by $a_{\mathrm{Erl}}(s, \rho)$, and the probability of being equal to s by $b_{\mathrm{Erl}}(s, \rho)$. Note that $b_{\mathrm{Erl}}(s, \rho)$ is equal to the famous Erlang B formula.

Theorem 7. *The stationary number of class-k customers in the system satisfies*

$$\mathrm{Erl}(m_k, \lambda_k/\mu_k) \leq_{\mathrm{st}} \bar{X}_{1,k} + \bar{X}_{2,k} \leq_{\mathrm{st}} \mathrm{Erl}(m_k + n, \lambda_k/\mu_k). \tag{37}$$

Especially, the stationary class-k mean number of customers is bounded by

$$a_{\mathrm{Erl}}(m_k, \lambda_k/\mu_k) \leq a_k \leq a_{\mathrm{Erl}}(m_k + n, \lambda_k/\mu_k), \tag{38}$$

the mean throughput by

$$\mu_k a_{\mathrm{Erl}}(m_k, \lambda_k/\mu_k) \leq \theta_k \leq \mu_k a_{\mathrm{Erl}}(m_k + n, \lambda_k/\mu_k), \tag{39}$$

and the blocking probability by

$$b_{\mathrm{Erl}}(m_k + n, \lambda_k/\mu_k) \leq b_k \leq b_{\mathrm{Erl}}(m_k, \lambda_k/\mu_k). \tag{40}$$

Proof. Let us consider a version of the process X started at $X(0) = 0$, and let $Z^{m_k}_{\lambda_k,\mu_k}$ be as in Theorem 5 and $Z^{m_k+n}_{\lambda_k,\mu_k}$ be as in Theorem 6, both started

at zero. Because all these processes are irreducible and positive recurrent, and because stochastic ordering is closed with respect to convergence in distribution [6], the inequalities (37) follow by taking $t \to \infty$ in (28) and (29).

The inequalities (38) follow by taking expectations, and the bounds (39) are a consequence of $\theta_k = \mu_k a_k$. In light of the conservation laws $\lambda_k(1-b_k) = \theta_k$ and $\lambda_k(1-b_{\mathrm{Erl}}) = \mu_k a_{\mathrm{Erl}}$, these bounds in turn imply (40). □

Figure 2 illustrates the bounds in (37) for a loss network with server configuration $m = (5,5)$ and $n = 5$, where $\lambda = (7.5, 7.5)$ and $\mu = (1, 1.3)$.

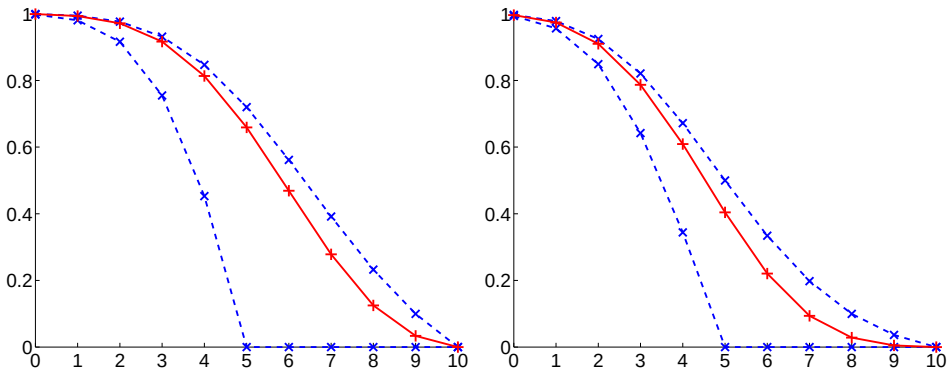

Figure 2: Complementary cumulative distribution functions of the stationary number of class-1 customers (left) and class-2 customers (right), plotted for the original system (solid line) and for the Erlang bounds in (37) (dotted lines).

Remark 4. The Erlang bounds (40) for the blocking probability are well-known in the literature (see for example [1]). The stochastic inequalities (37) can be viewed as extensions of these classical bounds.

5.2 Overall performance

Denote by $a = \mathrm{E}\,|\bar{X}|$ the stationary mean total number of customers, by $\theta = \sum_k \theta_k$ the stationary mean throughput, and by b the stationary overall blocking probability. Note that a may be viewed as the mean work throughput (net amount of work served by the system in unit time). We indicate by $a^{\mathrm{mp}}, \theta^{\mathrm{mp}}, b^{\mathrm{mp}}$ the corresponding quantities for a system with maximum packing.

Denote by $\mu^{\min}$ and $\mu^{\max}$ the vectors where all entries of μ are replaced by $\mu_{\min} = \min_k \mu_k$ and $\mu_{\max} = \max_k \mu_k$, respectively, and let $r_\mu = \mu_{\max}/\mu_{\min}$.

Moreover, let us denote by $C_{m,n}$ the set of server configurations where all layer-2 servers have been replaced by servers in layer 1, so that

$$C_{m,n} = \{m' \in \mathbb{Z}_+^K : m'_k \geq m_k \ \forall k \ \text{ and } \sum_k m'_k = \sum_k m_k + n\}.$$

Theorem 8. *The stationary total number of customers in the system satisfies*

$$|\bar{X}(m', 0, \lambda, \mu^{\max})| \leq_{\text{st}} |\bar{X}| \leq_{\text{st}} |\bar{X}^{\text{mp}}(m, n, \lambda, \mu^{\min})| \tag{41}$$

for all $m' \in C_{m,n}$. Especially, the stationary mean number of customers is bounded by

$$\max_{m' \in C_{m,n}} a(m', 0, \lambda, \mu^{\max}) \leq a \leq a^{\text{mp}}(m, n, \lambda, \mu^{\min}), \tag{42}$$

the mean throughput by

$$\max_{m' \in C_{m,n}} r_\mu^{-1} \theta(m', 0, \lambda, \mu^{\max}) \leq \theta \leq r_\mu \theta^{\text{mp}}(m, n, \lambda, \mu^{\min}), \tag{43}$$

and the overall blocking probability by

$$\begin{aligned} 1 - r_\mu(1 - b^{\text{mp}}(m, n, \lambda, \mu^{\min})) \\ \leq b \leq \min_{m' \in C_{m,n}} \left(1 - r_\mu^{-1}(1 - b(m', 0, \lambda, \mu^{\max}))\right). \end{aligned} \tag{44}$$

Remark 5. In the case where all service rates μ_k are equal, the bounds (43) and (44) can be written in a more natural form as

$$\begin{aligned} \max_{m' \in C_{m,n}} \theta(m', 0, \lambda, \mu) \leq \theta \leq \theta^{\text{mp}}(m, n, \lambda, \mu), \\ b^{\text{mp}}(m, n, \lambda, \mu) \leq b \leq \min_{m' \in C_{m,n}} b(m', 0, \lambda, \mu). \end{aligned}$$

Remark 6. The upper and lower bounds in (41), and hence the also the bounds in (42) – (44), are easy to compute numerically. The fast computation of the upper bound is explained in Remark 1. To compute the lower bound, observe that $|\bar{X}(m', 0, \lambda, \mu^{\max})|$ has the same distribution as $\sum_k \text{Erl}(m'_k, \lambda_k/\mu_{\max})$, where the terms in the sum are independent.

Proof of Theorem 8. Let X be the number of customers in the original system, let W be the number of customers in the system corresponding to the parameters $(m', 0, \lambda, \mu^{\max})$, and let Y be the number of customers in the maximum packing system with parameters $(m, n, \lambda, \mu^{\min})$. Assume that all

processes are started at zero initial state. Then Theorem 2 and Theorem 3 combined with Theorem 4 imply that

$$|W(t)| \leq_{\mathrm{st}} |X(t)| \leq_{\mathrm{st}} |Y(t)| \tag{45}$$

for all t. Because all of the above processes are irreducible and positive recurrent, and because stochastic ordering is closed with respect to convergence in distribution [6], taking $t \to \infty$ in (45) shows the validity of (41). The bounds in (42) follow by taking expectations, and the bounds in (43) from $\theta = \sum_k \mu_k a_k$. These bounds in turn imply (44), because of the conservation law $(\sum_k \lambda_k)(1-b) = \theta$. □

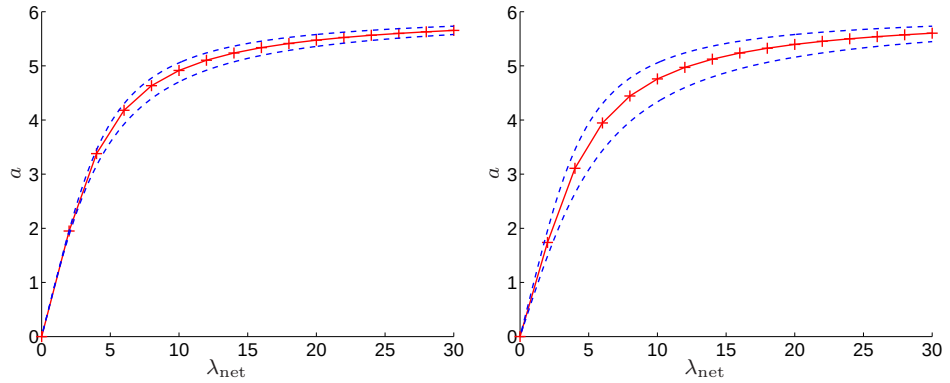

Figure 3: Stationary mean number of customers for $m = (2, 2)$, $n = 2$, and $\lambda = (\lambda_{\mathrm{net}}/2, \lambda_{\mathrm{net}}/2)$; where $\mu = (1, 1)$ on the left, and $\mu = (1, 1.3)$ on the right; plotted for the original system (solid line) and for the bounds in (42) (dotted lines).

Figure 3 illustrates the bounds (42) of the mean number of customers in a two-class system where the net arrival rate λ_{net} is varying. We see that the bounds are rather robust with respect to different values of the arrival rates. Figure 4 illustrates the same bounds for varying μ_2, showing that the accuracy of the bounds degrades rapidly as the difference of μ_2 and μ_1 grows. This loss of accuracy is an inevitable consequence of replacing μ by $\mu^{\min}$ and $\mu^{\max}$ in (42). Intuitively one might think that the upper bounds in (41) and (42) would hold without replacing μ by $\mu^{\min}$. Example 3 shows that this is not true in general. However, the right-hand side of (42) with μ in place of $\mu^{\min}$, though not generally an upper bound, appears to approximate well the original system for a wide range of system parameters, even for the

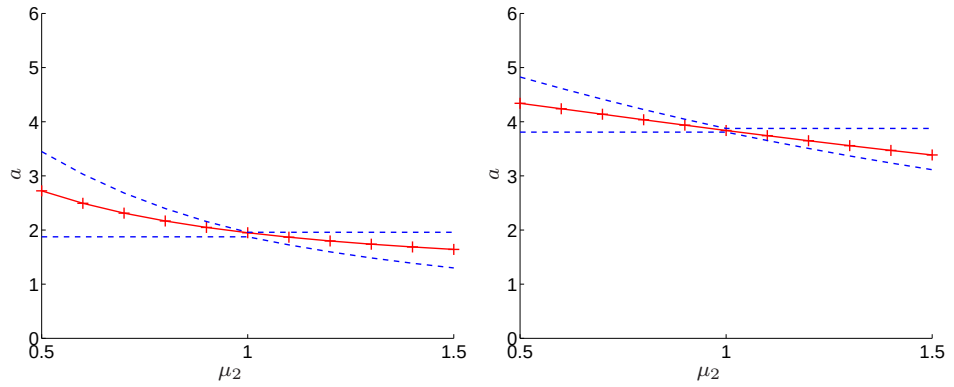

Figure 4: Stationary mean number of customers for $m = (2, 2)$, $n = 2$, $\mu_1 = 1$, and varying μ_2, where $\lambda = (1, 1)$ on the left and $\lambda = (1, 5)$ on the right; plotted for the original system (solid line) and for the bounds in (42) (dotted lines).

extreme choice of service rates of Example 3. The reason is that the actual repacking events in the maximum packing system occur relatively rarely in moderately loaded systems; see Kelly [7] for an insightful discussion of this phenomenon in the context of channel assignment in cellular radio networks.

Example 3. Consider a two-class loss network with server configuration $m = (1, 0)$ and $n = 2$. Assume $\lambda = (1, 1)$ and $\mu = (\frac{1}{5}, 10)$, so that the service rates differ from each other by a factor of 50. Table 1 lists numerically calculated values of the stationary mean number of customers (per class and total) for the original loss network and the modification with maximum packing. The fact $a(m, n, \lambda, \mu) > a^{\mathrm{mp}}(m, n, \lambda, \mu)$ illustrates that for this special choice of parameters, maximum packing does *not* increase the stationary mean number of customers in the system.

	Class 1	Class 2	Total
$a(m, n, \lambda, \mu)$	2.325657	0.038612	2.364269
$a^{\mathrm{mp}}(m, n, \lambda, \mu)$	2.317818	0.046344	2.364162
$a(m, n, \lambda, \mu^{\min})$	1.615744	0.997537	2.613281
$a^{\mathrm{mp}}(m, n, \lambda, \mu^{\min})$	1.474617	1.172442	2.647059

Table 1: Mean number of customers in a loss network with and without maximum packing.

Example 4 shows that replacing one layer-2 server by a layer-1 server

may not decrease the stationary mean number of customers, if not all service rates μ_k are equal. This shows that it is necessary to replace μ by $\mu^{\max}$ in order to achieve a lower bound in (41).

Example 4. Consider a two-class loss network with two different server configurations (i) $m = (0,0)$ and $n = 3$, and (ii) $m' = (1,0)$, $n' = 2$. Assume that λ and μ are as in Example 3. Numerically calculated values for the stationary mean number of customers (per class and total) given in Table 2. The fact $a(m,n,\lambda,\mu) < a(m',n',\lambda,\mu)$ illustrates that for this special choice of parameters, replacing one layer-2 server by a layer-1 server does *not* decrease the stationary mean number of customers.

	Class 1	Class 2	Total
$a(m,n,\lambda,\mu)$	2.317808	0.046356	2.364164
$a(m',n',\lambda,\mu)$	2.325657	0.038612	2.364269
$a(m,n,\lambda,\mu^{\max})$	0.099891	0.099891	0.199782
$a(m',n',\lambda,\mu^{\max})$	0.099906	0.099453	0.199359

Table 2: Mean number of customers in a loss network with two different server configurations.

6 Conclusions

Stochastic comparison techniques were developed for analyzing multiclass two-layer loss systems. First, assuming all service rates to be equal, we proved that maximum packing stochastically increases the total number of customers, and that moving a server from the second layer to the first has the opposite effect. The monotonicity of the system with respect to service rates was then used to extend the above conclusions to systems where the service rates may differ from each other. As a consequence, computationally fast upper and lower bounds for the performance of the system were derived.

The proofs of the main results (excluding Theorem 5) were based on coupling of continuous-time Markov processes, for which it was essential to assume that the service times are exponentially distributed. On the other hand, the stationary distributions of the processes acting as bounds in the main results, the maximum packing system and the Erlang loss system, are known to be insensitive to the service time distribution [8]. This remarkable feature calls for an extension of the comparison results to more general service time distributions. This is an important open problem for which we

believe that new probabilistic techniques are needed, because a purely deterministic sample path approach was found unsuitable (Examples 1 and 2).

The accuracy of the bounds was numerically studied for systems with small number of servers. The bounds for the per-class quantities appear not very accurate in general, though they may still be useful in conservative dimensioning of system resources. The bounds for the aggregate system quantities are much more accurate, especially when the mean service times across different customer classes do not vary too much. For highly variable mean service times, the accuracy degrades due to the need to modify the service time parameters in Theorem 8; however, if one uses the original μ in place of $\mu^{\min}$ in Theorem 8, the maximum packing system appears to approximate well the original system for a wide range of system parameters (though not anymore an upper bound in general, see Example 3). The accurate numerical evaluation of the system becomes difficult when the number of servers is large, because of the rapid growth of the state space [11]. An interesting future problem is to asymptotically study the sharpness of the bounds for large systems using scaling and renormalization techniques.

Acknowledgments

We gratefully acknowledge helpful discussions with Sem Borst and sharp remarks by anonymous referees. The main part of this research was carried out at Centrum voor Wiskunde en Informatica and Eindhoven University of Technology. The research has been supported by the Dutch BSIK/BRICKS PDC2.1 project, Helsingin Sanomat Foundation, and the Academy of Finland.

References

[1] Borst, S. and Whiting, P. A. (2000). Achievable performance of dynamic channel assignment schemes under varying reuse constraints. *IEEE T. Veh. Technol.*, 49(4):1248–1264.

[2] Everitt, D. E. and Macfadyen, N. W. (1983). Analysis of multicellular mobile radiotelephone systems with loss. *Brit. Telecom Technol. J.*, 1(2):37–45.

[3] Franx, G. J., Koole, G., and Pot, A. (2006). Approximating multi-skill blocking systems by hyperexponential decomposition. *Perform. Evaluation*, 630:799–824.

[4] Hordijk, A. and Ridder, A. (1987). Stochastic inequalities for an overflow model. *J. Appl. Probab.*, 24:696–708.

[5] Kallenberg, O. (2002). *Foundations of Modern Probability*. Springer, second edition.

[6] Kamae, T., Krengel, U., and O'Brien, G. L. (1977). Stochastic inequalities on partially ordered spaces. *Ann. Probab.*, 5(6):899–912.

[7] Kelly, F. P. (1985). Stochastic models of computer communication systems. *J. Roy. Stat. Soc. B*, 47(3):379–395.

[8] Kelly, F. P. (1991). Loss networks. *Ann. Appl. Probab.*, 1:319–378.

[9] Last, G. and Brandt, A. (1995). *Marked Point Processes on the Real Line: The Dynamical Approach*. Springer.

[10] Lindvall, T. (1999). On Strassen's theorem on stochastic domination. *Electron. Commun. Prob.*, 4:51–59.

[11] Louth, G., Mitzenmacher, M., and Kelly, F. P. (1994). Computational complexity of loss networks. *Theor. Comp. Sc.*, 125:45–59.

[12] Massey, W. A. (1987). Stochastic orderings for Markov processes on partially ordered spaces. *Math. Oper. Res.*, 12(2):350–367.

[13] Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Wiley.

[14] Nain, P. (1990). Qualitative properties of the Erlang blocking model with heterogeneous user requirements. *Queueing Syst.*, 6:189–206.

[15] Smith, D. R. and Whitt, W. (1981). Resource sharing for efficiency in traffic systems. *Bell System Tech. J.*, 60(1):39–55.

[16] Thorisson, H. (2000). *Coupling, Stationarity, and Regeneration*. Springer.

[17] Whitt, W. (1981). Comparing counting processes and queues. *Adv. Appl. Probab.*, 13(1):207–220.

[18] Wolff, R. W. (1989). *Stochastic Modeling and the Theory of Queues*. Prentice Hall.